ESTIMACIÓN DE PRECIOS DE ALQUILER DE VIVIENDAS MEDIANTE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

R. T. Mora García, M. F. Céspedes López, V. R. Pérez Sánchez, J. C. Pérez Sánchez

Universidad de Alicante, San Vicente Del Raspeig, España

RESUMEN

Este trabajo se enmarca en un contexto de creciente interés por la aplicación de técnicas de inteligencia artificial (IA) en el mercado inmobiliario, especialmente en un momento en el que España se encuentra con importantes incrementos del precio de alquiler y compra de vivienda. La investigación aborda la aplicación de algoritmos de aprendizaje automático para estimar precios de alquiler y desarrollar una plataforma web abierta que acerque estas tecnologías tanto a usuarios individuales como a profesionales del sector. El objetivo principal consiste en diseñar una metodología de entrenamiento, optimización e interpretación de modelos predictivos orientados al cálculo automatizado del precio de alquiler de viviendas multifamiliares en la Comunidad Valenciana. Este objetivo se materializa en la creación de una aplicación web que permita realizar estimaciones actualizadas y fácilmente interpretables tanto para usuarios especializados como no especializados. La base de datos utilizada ha sido obtenida de portales inmobiliarios, de los cuales se han descargado mensualmente datos durante dos años (2024–2025), incluyendo precios de oferta, atributos de la vivienda y del edificio, su ubicación espacial (coordenadas geográficas) y un identificador temporal. Tras un exhaustivo proceso de limpieza, depuración y eliminación de duplicados, se conformó un conjunto de datos de corte transversal agrupado (*pooled cross-section*). La metodología empleada se ajustó a las fases establecidas del proceso de aprendizaje automático: preparación y análisis exploratorio de datos, ingeniería de características, entrenamiento de modelos, optimización de hiperparámetros, evaluación, interpretación y despliegue. Se compararon varios algoritmos de aprendizaje supervisado basados en conjuntos (*ensemble learning*) como el *boosting* (GBR, XGBM, LGBM) y el *bagging* (RF, ET), tomando como línea base la regresión lineal por mínimos cuadrados ordinarios. La optimización se realizó mediante estrategias de búsqueda aleatoria y bayesiana con validación cruzada, utilizando particiones agrupadas por el identificador del inmueble. En la evaluación de los modelos se emplearon métricas de error (MAE, MSE, RMSE) y bondad de ajuste (R2) aplicadas a conjuntos de entrenamiento y prueba. Los resultados demuestran que los modelos basados en técnicas de *boosting* ofrecen un mejor desempeño predictivo y una mayor estabilidad frente a alternativas tradicionales y de *bagging*. Además, muestran una capacidad de generalización adecuada y un buen equilibrio entre precisión y eficiencia computacional. La interpretación del modelo mediante valores de Shapley (SHAP) y análisis de importancia por permutación revela que las variables más influyentes son la ubicación geográfica, la renta neta y las características intrínsecas del inmueble como superficie, número de baños y dormitorios. La presencia de ascensor se confirma como un factor

de impacto marginal. El estudio culmina con la implementación de una aplicación web de acceso abierto, desarrollada en Python y Streamlit, que permite introducir las características de un inmueble, indicar su localización y obtener una estimación del precio de oferta acompañada de explicaciones gráficas y textuales. La herramienta democratiza el acceso a la valoración automatizada, aportando transparencia y utilidad práctica a ciudadanos y profesionales.

PALABRAS CLAVE: precio de la vivienda, valoración masiva, aprendizaje automático, hiperparámetros, comunidad valenciana.

1. INTRODUCCIÓN

Durante la última década, la inteligencia artificial (IA) ha pasado de ser un ámbito eminentemente académico para convertirse en una tecnología transversal con impactos tangibles en la vida cotidiana. Sus aplicaciones abarcan desde asistentes conversacionales y sistemas de recomendación hasta visión por computador, procesamiento del lenguaje natural, diagnóstico asistido, conducción autónoma o predicción de fenómenos complejos. Esta expansión se explica, principalmente, por la disponibilidad de grandes volúmenes de datos digitales, el abaratamiento de la capacidad de cómputo y el avance de algoritmos capaces de aproximar relaciones no lineales entre variables de manera eficiente.

Sin embargo, una limitación crítica de muchos algoritmos de IA contemporáneos es su opacidad: ofrecen predicciones precisas, pero no necesariamente justifican de forma comprensible cómo y por qué alcanzan un resultado. Este rasgo, habitualmente descrito como caja negra, adquiere especial relevancia cuando los modelos se aplican a decisiones con efectos económicos o sociales directos. La situación se acentúa en soluciones desarrolladas por empresas privadas, donde la explicabilidad puede quedar restringida por razones de propiedad intelectual, secretos comerciales o estrategias competitivas. En este marco, la incorporación de enfoques de IA explicable (*explainable AI*, XAI) resulta clave para dotar de trazabilidad a las predicciones, facilitar su auditoría, detectar sesgos y aumentar la confianza de los usuarios [1].

La IA ha comenzado a desempeñar un papel significativo en el mercado inmobiliario, introduciendo innovaciones que mejoran la eficiencia, la personalización y la toma de decisiones de los usuarios. Ejemplo de ellos son las plataformas inmobiliarias que utilizan la IA en la búsqueda y recomendación de propiedades; asistentes de visitas virtuales a inmuebles; el análisis predictivo del mercado inmobiliario; la evaluación del riesgo, el crédito o el fraude financiero, entre otras aplicaciones.

En España, este proceso tecnológico se superpone a un contexto de tensión en los precios de la vivienda, tanto en compra como en alquiler, que dificulta el acceso residencial y amplifica la incertidumbre de los hogares. La evolución no ha sido homogénea territorialmente: el encarecimiento tiende a concentrarse en grandes áreas urbanas y enclaves turísticos, donde la presión de demanda, la limitada oferta disponible y determinados usos alternativos del parque residencial intensifican el problema. A ello se suman factores que afectan al coste de financiación y a la producción inmobiliaria —como el aumento de tipos de interés, la inflación y el encarecimiento de materiales—, así como cambios regulatorios recientes que inciden en los incentivos y expectativas de arrendadores y

demandantes. En este escenario, disponer de referencias cuantitativas consistentes y transparentes sobre precios de oferta puede contribuir a decisiones más informadas.

Pese a ello, la ciudadanía suele enfrentarse al mercado inmobiliario con asimetrías de información, con un conocimiento limitado sobre dinámicas locales, falta de criterios comparables y dificultades para interpretar datos heterogéneos. En paralelo, existe una brecha en el uso efectivo de herramientas digitales avanzadas para comprender y contrastar valores de mercado. Desde esta perspectiva, resulta pertinente promover capacidades de "ciudadanía digital" orientadas a un uso crítico y responsable de soluciones basadas en datos, especialmente cuando dichas soluciones influyen en decisiones patrimoniales de alto impacto. En consecuencia, facilitar instrumentos accesibles y explicables, no solo aporta utilidad práctica inmediata, sino que también contribuye a la alfabetización digital aplicada a un problema socialmente relevante.

A partir de estas consideraciones, los autores de esta investigación consideran que proporcionar estimaciones transparentes y actualizadas del precio de alquiler y venta, basadas en datos reales de oferta y en modelos interpretables, puede apoyar tanto a demandantes (hogares, jóvenes, colectivos con menor renta) como a agentes profesionales. Entre estos últimos se incluyen intermediarios y comercializadores, tasadores inmobiliarios, técnicos de la administración, promotores y pequeños o medianos inversores, que pueden beneficiarse de esta nueva fuente de información y tecnología sobre el mercado inmobiliario propuesta en esta investigación.

El objetivo principal del trabajo es desarrollar una metodología para el entrenamiento y la optimización de algoritmos de aprendizaje automático (*Machine Learning* en inglés, ML) orientados a la estimación del precio de oferta del alquiler de viviendas multifamiliares en la Comunidad Valenciana. Este objetivo se materializa en una aplicación web de acceso abierto diseñada para ofrecer estimaciones interpretables en municipios grandes y medianos, reduciendo barreras de uso para personas con competencias tecnológicas limitadas. La contribución del estudio es doble: 1) una estrategia metodológica reproducible para el tratamiento de datos procedentes de portales inmobiliarios y el ajuste de modelos con validación robusta; y 2) un despliegue práctico que traslada el resultado científico a una herramienta utilizable, con énfasis en la explicación de la predicción.

La estimación automatizada del precio de la vivienda ha recibido una atención creciente en la literatura científica, impulsada por la disponibilidad de datos estructurados y georreferenciados, y por la consolidación de métodos capaces de capturar relaciones complejas entre atributos y valor de mercado. Un primer eje de investigación compara el rendimiento de modelos econométricos tradicionales —en particular, los modelos hedónicos— con algoritmos de ML, evaluando el compromiso entre interpretabilidad económica y capacidad predictiva [2]. La evidencia acumulada sugiere que, aunque los modelos clásicos son útiles para interpretar elasticidades y determinantes del precio, suelen mostrar limitaciones en entornos con alta dimensionalidad, no linealidades e interacciones complejas, donde los algoritmos de ML tienden a mejorar el desempeño predictivo [3-13].

Un segundo eje se centra en determinar qué familias de algoritmos resultan más adecuadas según el contexto espacial y temporal, comparando métodos como *Random Forest*, *Gradient Boosting*, *XGBoost*, *LightGBM* o redes neuronales, y utilizando métricas de error (MAE, MSE, RMSE) y ajuste (R^2) para establecer diferencias de rendimiento [2, 14-18]. Un gran número de estos trabajos reportan la

superioridad de enfoques basados en *boosting*, especialmente por su capacidad para modelar no linealidades, gestionar interacciones sin especificación funcional previa y mantener un equilibrio competitivo entre precisión y eficiencia computacional.

Este artículo se organiza como sigue. La Sección 2 presenta la metodología y la construcción de la base de datos. La Sección 3 expone los resultados del entrenamiento, la optimización y la validación comparada de modelos, junto con su interpretación mediante técnicas XAI. Por último, la Sección 4 sintetiza las conclusiones.

2. MATERIALES Y MÉTODO

2.1. Fuentes de información y base de datos

La información utilizada en este estudio procede de un portal inmobiliario del que se extrajeron precios de oferta de alquiler y un conjunto amplio de atributos descriptivos del inmueble. La recolección se realizó con periodicidad mensual durante dos años (2024–2025), lo que permite capturar tanto la heterogeneidad espacial del mercado como su evolución temporal a lo largo de distintos ciclos.

Cada registro incluye (i) el precio ofertado del alquiler; (ii) características intrínsecas de la vivienda (p. e., tipología, superficie construida, número de dormitorios, baños y aseos, equipamientos); (iii) atributos del edificio (p. e., ascensor, garaje, trastero, piscina, terraza); y (iv) su localización geográfica, codificada mediante coordenadas en grados decimales (datum WGS84). Esta estructura permite trabajar con modelos capaces de integrar simultáneamente información estructural y espacial.

A partir de la descarga inicial se aplicó un protocolo de depuración en varias etapas. Primero, mediante un control de calidad y consistencia de los datos, realizando una revisión de formatos y detección de valores incompatibles o inconsistencias con las estructura de los datos. Segundo, mediante la gestión de valores ausentes, descartando observaciones sin información relevante para la predicción (inmuebles sin precio, sin coordenadas o sin atributos esenciales). Tercero, seleccionando variables relevantes, y eliminando otras con muy baja variabilidad (poca capacidad discriminante) o con un porcentaje elevado de valores faltantes, con el fin de reducir ruido y mejorar la estabilidad del entrenamiento. Por último, se procedió a la eliminación de registros duplicados, identificando y suprimiendo anuncios redundantes con atributos idénticos, para evitar sobrerrepresentación artificial de determinadas viviendas.

Para representar el componente temporal se incorporó una variable categórica año y trimestre, que identifica el trimestre en el que el anuncio estuvo activo o fue capturado. Esta codificación permite modelar cambios en las tendencias del mercado sin imponer una forma funcional rígida. Finalmente, el conjunto se estructuró como una muestra de corte transversal agrupado (*pooled cross-section*), donde las observaciones corresponden a anuncios en distintos momentos del tiempo, sin asumir necesariamente una estructura de panel equilibrado.

Tras la depuración, la base de datos de viviendas multifamiliares en alquiler quedó compuesta por 46.900 inmuebles únicos, la serie completa de capturas trimestrales generó 84.778 observaciones para el periodo 2024–2025. La Tabla 1 resume las variables utilizadas, agrupadas por categoría y con indicación del tipo de codificación aplicado.

Tabla 1. Características del conjunto de variables utilizadas

Categoría	Característica	Tipo de valor	Descripción
Características de la vivienda	tipología	Categórico	Variables ficticias que identifican la tipología de la vivienda (piso, ático, dúplex/tríplex, estudio/loft y planta baja)
	superficie	Numérico	Superficie construida de la vivienda (m^2)
	dormitorios	Numérico	Número de dormitorios de la vivienda
	baños	Numérico	Número de baños de la vivienda
	aseos	Numérico	Número de aseos de la vivienda
	aire_acond	Dicotómico	Disponibilidad de aire acondicionado
	calefaccion	Dicotómico	Disponibilidad de calefacción
	terraza	Dicotómico	Disponibilidad de terraza descubierta
	obra_nueva	Dicotómico	Indica si la vivienda es de obra nueva
Características del edificio	ascensor	Dicotómico	Disponibilidad de ascensor
	garaje	Dicotómico	Disponibilidad de plaza de garaje
	trastero	Dicotómico	Disponibilidad de trastero
	piscina	Dicotómico	Disponibilidad de piscina
Características de ubicación	longitud	Numérico	Coordenadas geográficas de la ubicación espacial (en grados decimales), datum WGS84
	latitud	Numérico	
	renta_neta	Numérico	Renta neta media por hogar en miles de euros [19], del año 2020 obtenidas a partir de las secciones censales
	provincia	Categórico	Variables ficticias que identifican la provincia donde se ubica cada inmueble
Características temporales	trimestre	Categórico	Variables ficticias para modelar el factor tiempo (8 trimestres, desde 2024T1 hasta 2025T4)

Variable dependiente	precio	Numérico	El precio de alquiler ofertado de viviendas multifamiliares en euros por mes (€/mes)

2.2. Metodología

La investigación se apoya en un enfoque de aprendizaje supervisado para regresión y, en particular, en modelos de aprendizaje conjunto (*ensemble learning*) capaces de capturar relaciones no lineales e interacciones complejas entre atributos estructurales, localización y tiempo. Se evaluaron dos familias principales: 1) Modelos basados en *boosting*, como *Gradient Boosting Regressor* (GBR), *Extreme Gradient Boosting* (XGBoost/XGBM) y *Light Gradient Boosting Machine* (LightGBM/LGBM); y 2) Modelos basados en *bagging*, como *Random Forest* (RF) y *Extra Trees Regressor* (ET).

La implementación se realizó en Python, empleando herramientas estándar del ecosistema científico: pandas y NumPy para tratamiento de datos; scikit-learn como marco principal para modelado, validación y métricas; xgboost y lightgbm para los modelos de *boosting* de alto rendimiento; y scikit-optimize para la optimización de hiperparámetros. La generación de figuras se abordó con matplotlib (y librerías auxiliares de visualización), y la interpretación del modelo se realizó mediante SHAP y técnicas de importancia por permutación.

El flujo metodológico seguido se estructura en siete etapas, alineadas con las prácticas habituales de *machine learning* aplicado:

1) Preparación y exploración de datos. Se efectuó un análisis exploratorio inicial para caracterizar distribuciones, detectar inconsistencias y revisar rangos admisibles. Posteriormente se aplicaron tareas de depuración (tratamiento de ausencias en variables críticas, control de valores extremos y estandarización de formatos), de acuerdo con el protocolo descrito en la Sección 2.1.
2) Preprocesamiento e ingeniería de variables. Las variables numéricas se incorporaron directamente al modelo, mientras que las variables categóricas (p. e., tipología, provincia y trimestre) se codificaron mediante variables categóricas. En esta etapa se revisó la colinealidad y la redundancia de información, priorizando un conjunto final de predictores que equilibrara capacidad explicativa y parsimonia.
3) Diseño de particiones y validación. Para evaluar la capacidad de generalización y evitar fugas de información derivadas de la reaparición de un mismo inmueble en distintas capturas temporales, las particiones de entrenamiento y prueba se realizaron de forma agrupada por identificador de inmueble (*grouped split*). De este modo, los registros de una misma vivienda quedan contenidos en un único subconjunto, reduciendo el riesgo de sobrestimación en la evaluación del rendimiento.
4) Entrenamiento de modelos candidatos y línea base. Se entrenaron los cinco algoritmos *ensemble* mencionados, comparando su rendimiento con un modelo de referencia de complejidad inferior (línea base) para contextualizar las ganancias obtenidas. Esta comparación permite valorar el incremento de precisión frente al aumento de complejidad y coste computacional.
5) Optimización de hiperparámetros. Sobre los modelos con mejor desempeño preliminar se aplicaron estrategias de ajuste de hiperparámetros mediante validación cruzada, combinando

búsqueda aleatoria y optimización bayesiana. Este procedimiento permite explorar el espacio de configuraciones de manera eficiente y reducir el riesgo de seleccionar hiperparámetros sobreajustados a una partición concreta.

6) Evaluación y selección del modelo. El desempeño se cuantificó con métricas complementarias de error y ajuste, incluyendo MAE, MSE, RMSE y R^2, reportadas sobre conjuntos de entrenamiento y prueba. Adicionalmente, se analizó la estabilidad del rendimiento y la presencia de sobreajuste comparando resultados entre particiones y evaluando la consistencia del error.
7) Interpretabilidad y despliegue. Una vez seleccionado el modelo final, se aplicaron técnicas de interpretación globales y locales para caracterizar la contribución de cada variable a la predicción, combinando importancia por permutación y valores de Shapley (SHAP). Finalmente, el modelo se integró en una aplicación web para su uso operativo, permitiendo introducir las características de un inmueble y obtener una estimación acompañada de explicaciones visuales orientadas a la trazabilidad del resultado.

3. RESULTADOS Y DISCUSIÓN

3.1. Entrenamiento y optimización de los modelos

La fase de optimización de hiperparámetros se planteó con un objetivo operativo claro: mejorar la capacidad de ajuste y reducir el error de predicción sin comprometer la generalización. Para ello, el procedimiento se ejecutó exclusivamente sobre la partición de entrenamiento, correspondiente al 70% del conjunto total. Sobre este subconjunto se aplicó validación cruzada de cinco pliegues, generando iterativamente pares CV-entrenamiento / CV-validación para estimar el rendimiento esperado del modelo bajo distintas configuraciones.

Con el fin de evitar fugas de información derivadas de la reaparición temporal de anuncios del mismo inmueble, todas las particiones —tanto el corte entrenamiento/prueba como las divisiones internas de validación cruzada— se realizaron de forma agrupada por identificador de vivienda, garantizando que los registros asociados a un inmueble quedaran contenidos en un único subconjunto. Este criterio se implementó mediante el esquema de separación por grupos disponible en scikit-learn (método *GroupShuffleSplit*), utilizando como variable de agrupación el identificador del inmueble.

Una vez definida la canalización de entrenamiento, cada algoritmo se optimizó mediante dos estrategias complementarias de búsqueda de hiperparámetros: búsqueda aleatoria y optimización bayesiana, fijando en ambos casos un máximo de 20 iteraciones. Este límite se adoptó como compromiso entre exhaustividad y coste computacional, dado que determinados modelos —especialmente los basados en *bagging* como Random Forest (RF) y Extra Trees (ET)— presentan tiempos de entrenamiento elevados cuando se combinan con validación cruzada y exploración del espacio de hiperparámetros.

La Figura 1 resume el rendimiento obtenido en la fase de búsqueda de hiperparámetros, representando mediante un diagrama de cajas la distribución de los valores de R^2 alcanzados en los pliegues de CV-entrenamiento y CV-validación, considerando ambas estrategias de búsqueda para cada algoritmo. Esta visualización permite comparar no solo el valor central del desempeño, sino también su dispersión, aportando una lectura de estabilidad y sensibilidad al proceso de optimización.

Los resultados muestran un patrón consistente: los algoritmos basados en *boosting* ofrecieron el mejor desempeño durante la optimización, con LGBM como alternativa más competitiva, seguido de GBR y XGBM. En cambio, los modelos basados en *bagging* (RF y ET) no alcanzaron valores de R^2 comparables en CV-validación, evidenciando una menor capacidad para capturar las no linealidades e interacciones presentes en los datos en el contexto evaluado. En conjunto, estos hallazgos justifican que la selección del modelo final se concentre en la familia de *boosting*, al combinar mejores niveles de ajuste con una variabilidad más controlada en los pliegues de validación.

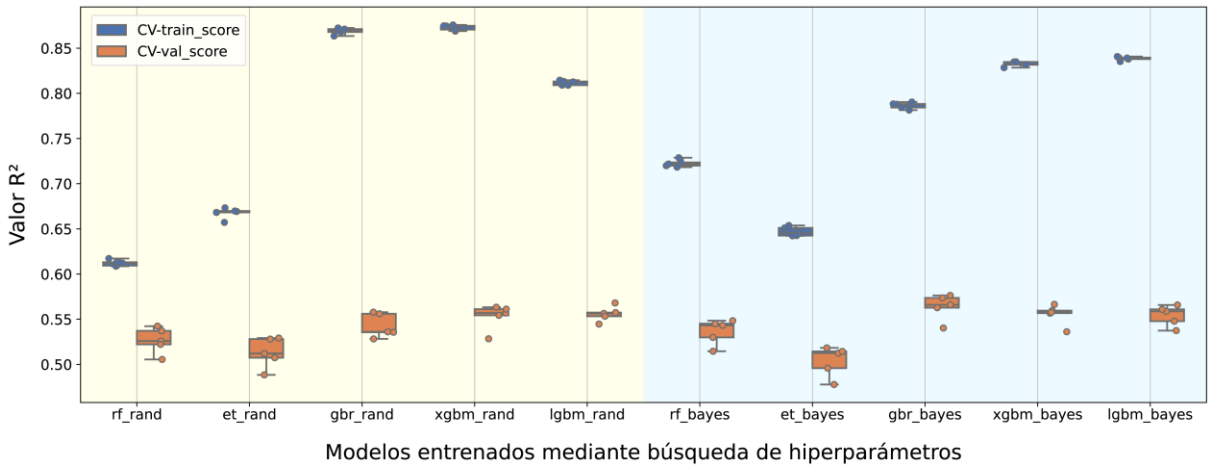


Figura 1. Resultados de rendimiento del ajuste de hiperparámetros mediante validación cruzada con estrategias de búsqueda aleatoria y bayesiana

3.2. Evaluación y selección de los modelos

En esta etapa se evaluó el desempeño predictivo de los algoritmos candidatos sobre un subconjunto de prueba independiente, equivalente al 30% del conjunto de datos original. La selección del modelo se basó en las métricas de error y de bondad de ajuste definidas en la metodología (MAE, MSE, RMSE y R^2), con el objetivo de identificar los algoritmos que maximizan la precisión sin comprometer su capacidad de generalización. Una vez determinada, en la fase previa, la configuración óptima de hiperparámetros para cada algoritmo, se procedió al entrenamiento final sobre el subconjunto de entrenamiento (70%) y a la extracción de métricas tanto en entrenamiento como en prueba.

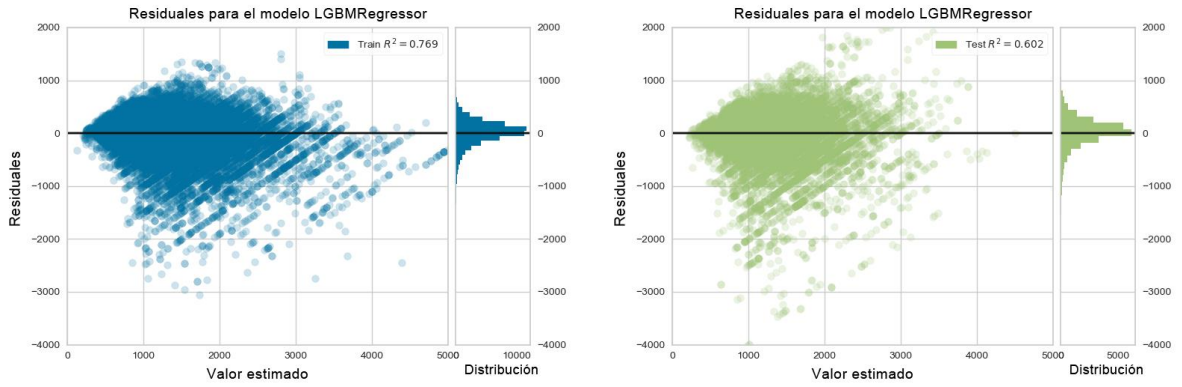
En la Tabla 2 se muestran métricas basadas en R^2 (valores más altos implica mejor rendimiento) y en error (MAE, MSE y RMSE) calculadas sobre el conjunto de prueba. Los valores de R^2 más altos corresponden a los modelos basados en *boosting*, destacando LGBM. Dado que las métricas de error penalizan directamente la desviación entre predicción y valor observado, valores inferiores implican mejor rendimiento. Los resultados corroboran el patrón observado con R^2: los modelos basados en *boosting* (LGBM, GBR y XGBM) concentran los menores errores, con valores muy próximos entre sí y claramente mejores que los obtenidos por RF, ET y, especialmente, la regresión lineal.

Tabla 2. Resultados del rendimiento de los algoritmos entrenados en el subconjunto de prueba

	R^2	MAE	MSE	RMSE
Linear Regression	0,3855	328,1	223.487,4	472,7
Random Forest Regressor	0,5444	261,8	165.683,2	407,0
Extra Trees Regressor	0,4865	288,6	186.746,6	432,1
Gradient Boosting Regressor	0,5990	237,6	145.836,6	381,9
Extreme Gradient Boosting	0,5969	240,0	146.590,1	382,9
Light Gradient Boosting Machine	**0,6023**	**235,6**	**144.649,0**	**380,3**

* R^2 Coeficiente de determinación; MAE Mean Absolute Error; MSE Mean Square Error; RMSE Root Mean Squared Error.

El análisis cualitativo mediante gráficos de residuales aporta evidencia adicional sobre la generalización. En la Figura 2 se representa el comportamiento del modelo LGBM en entrenamiento y prueba. La nube de residuos aparece dispersa y aproximadamente centrada en cero, sin patrones sistemáticos evidentes, lo que sugiere que el modelo captura adecuadamente la relación funcional entre predictores y precio. Además, la similitud entre las distribuciones de residuos en ambos conjuntos apunta a una incidencia limitada de sobreajuste.



(a) **(b)**

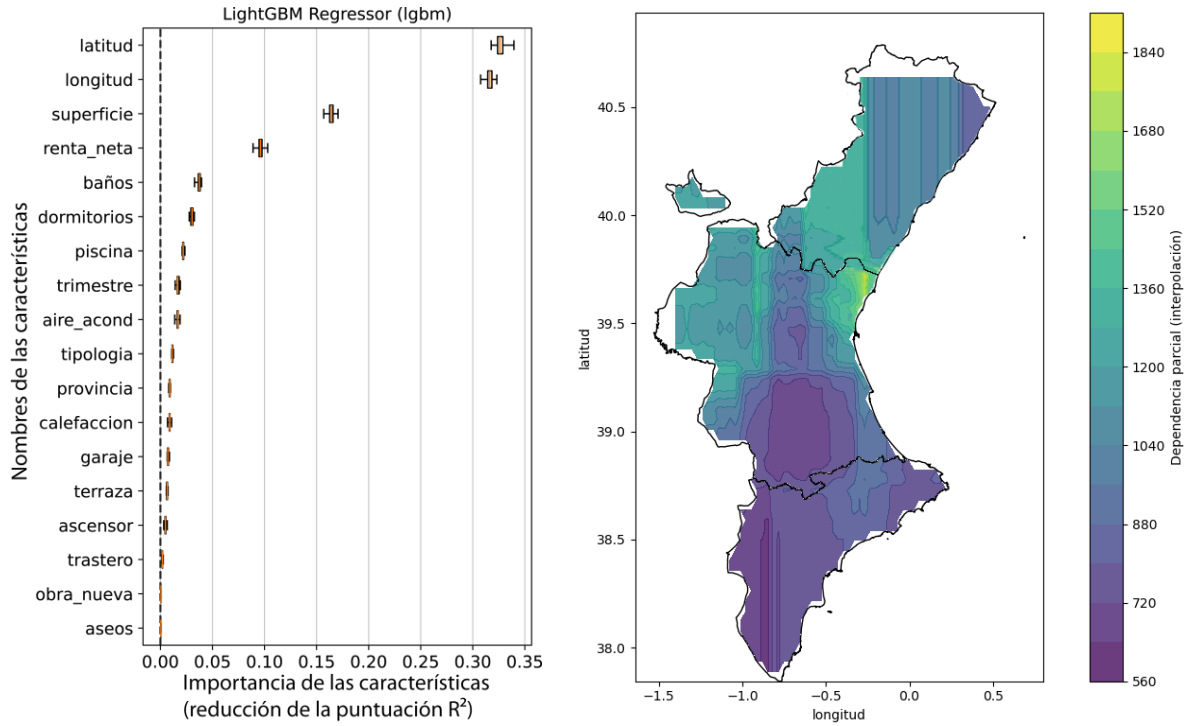
Figura 2. Gráficos de residuales del algoritmo entrenado LGBM**: (a)** Subconjunto de datos de entrenamiento. **(b)** Subconjunto de datos de prueba

Desde una perspectiva operativa, el coste computacional también es un criterio relevante en entornos de producción y actualización periódica. En las pruebas realizadas, XGBM y LGBM mostraron los

tiempos de entrenamiento más competitivos, siendo XGBM el más rápido a lo largo de los tamaños muestrales considerados. En contraste, RF y ET presentaron tiempos significativamente más elevados, lo que limita su escalabilidad cuando se trabaja con bases de datos extensas o cuando se requiere reentrenamiento frecuente.

3.3. Interpretación de los modelos

Con el fin de aportar transparencia al proceso predictivo y facilitar una lectura sustantiva de los resultados, la interpretación se abordó desde dos perspectivas complementarias: (i) importancia global de variables y efectos promedio, y (ii) explicación local para observaciones individuales. Para ello se emplearon, respectivamente, importancia por permutación, gráficos de dependencia parcial (PDP) y valores de Shapley (SHAP).



(a) **(b)**

Figura 3. Gráficos para la interpretación del modelo LGBM**: (a)** Importancia relativa de las características más relevantes según el algoritmo LGBM. **(b)** Gráfico de dependencia parcial bidireccional para las coordenadas longitud y latitud con el algoritmo LGBM

En la Figura 3a se muestran los resultados de la importancia de características mediante permutación para el modelo con mejor desempeño (LGBM). Las variables asociadas a la ubicación geográfica ocupan posiciones dominantes, destacando especialmente la latitud. Para profundizar en la importancia de la localización, la Figura 3b presenta el Gráfico de Dependencia Parcial (PDP) bidireccionales correspondientes a la longitud y latitud, para el modelo LGBM. Este resultado sugiere que existe una variación espacial entre provincias marcada del precio de oferta a lo largo del eje norte–sur de la

Comunidad Valenciana, así como en el eje este–oeste coincidente con la concentración de demanda y valor en áreas litorales y su progresiva reducción hacia el interior. Entre las características intrínsecas del inmueble, la superficie construida, el número de baños y el número de dormitorios se consolidan como determinantes principales. La renta media por hogar también ocupa una posición relevante en las características que más influyen en la determinación del precio de alquiler.

En lo que respecta a la interpretación local del modelo, se han utilizado los valores de Shapley (SHAP), una técnica basada en la teoría de juegos cooperativos que permite descomponer el resultado de una predicción en las contribuciones individuales de cada característica. En la Figura 4, se presenta un gráfico tipo cascada (*waterfall*) que muestra la estimación SHAP para un caso específico: un inmueble ofertado en alquiler en la ciudad de Alicante durante el cuarto trimestre de 2024. En este gráfico, el valor E[f(x)] representa la predicción promedio del modelo sobre toda la muestra (821 €), que actúa como valor base. A partir de este valor, se suman o restan las contribuciones de cada característica hasta alcanzar la predicción final del modelo, f(x)=1.111 €. Las características que incrementan el precio aparecen en rojo, mientras que aquellas que lo reducen se representan en azul. En este ejemplo, la variable trimestre es la que más impulsa el precio al alza, con una contribución de +226 €, mientras que la ausencia de aire acondicionado tiene un impacto negativo de –74 € en la estimación final. Este tipo de explicaciones aporta una visión transparente y comprensible del funcionamiento del modelo a nivel individual, facilitando su aplicación práctica y la confianza en sus predicciones.

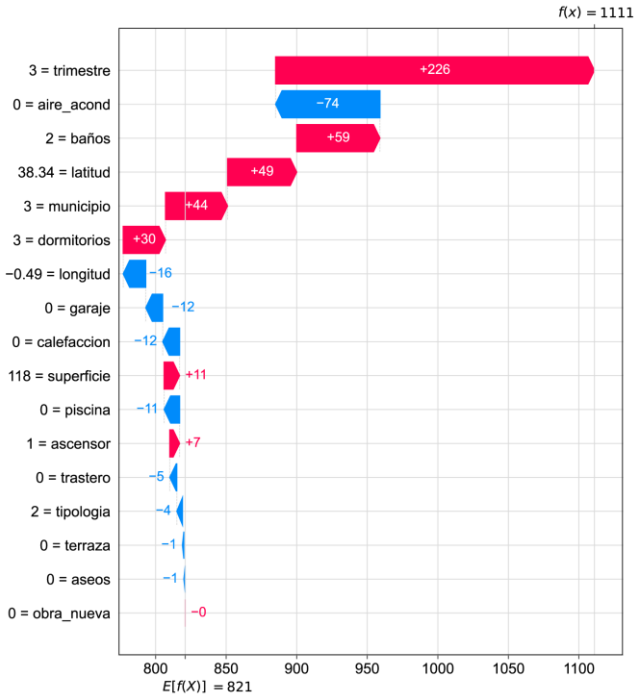


Figura 4. Gráfico de cascada para una observación de la base de datos de alquiler, estimación realizada utilizando el algoritmo LGBM.

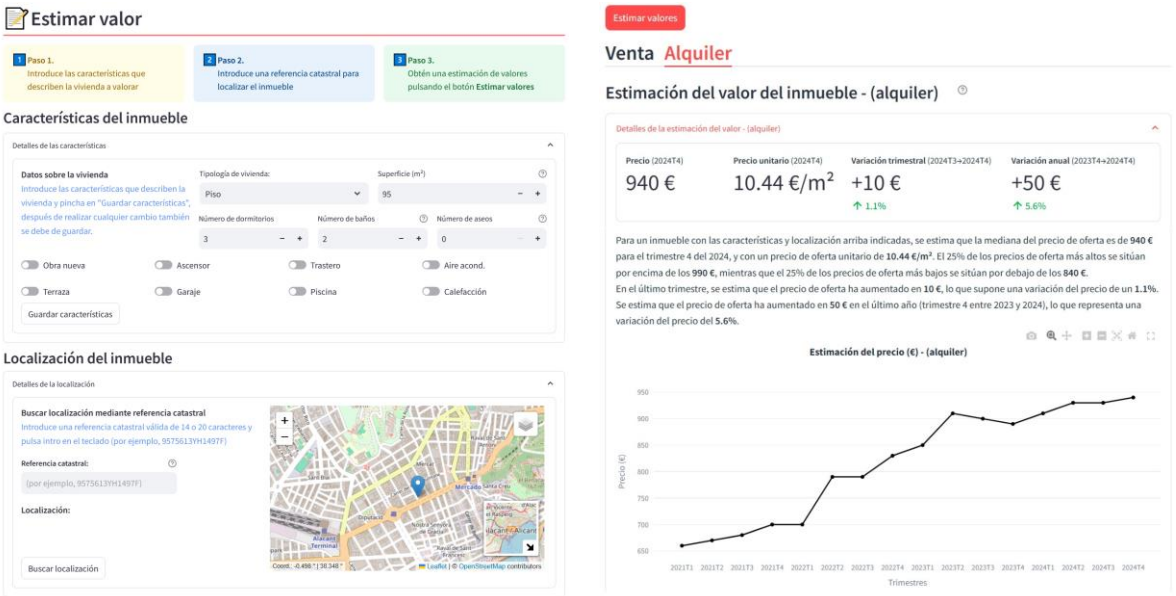
3.4. Despliegue de los modelos

En la fase final del estudio se desarrolló una aplicación web orientada a poner en producción el modelo predictivo y permitir la obtención de estimaciones a partir de la información introducida manualmente por el usuario (web https://preciosdevivienda.streamlit.app/). La herramienta se ofrece en acceso

abierto con un doble propósito: (i) facilitar su utilización por parte de la ciudadanía y (ii) proporcionar a profesionales del sector inmobiliario e inversores un entorno sencillo para realizar simulaciones y contrastar escenarios de precio de oferta del alquiler.

La interfaz se estructura en tres bloques funcionales: (1) características del inmueble, (2) localización geográfica (ver Figura 5a), y (3) resultados de la estimación (Figura 5b). En el primer bloque el usuario introduce los atributos del inmueble necesarios para la predicción (tipología, superficie construida, dormitorios, baños, aseos y equipamientos). En el segundo bloque se solicita una referencia catastral válida (de 14 o 20 caracteres). El sistema realiza una validación automática de la referencia y, cuando la verificación es satisfactoria, devuelve una dirección aproximada junto con las coordenadas geográficas asociadas, que se incorporan como entrada espacial del modelo.

En el bloque de resultados, la aplicación presenta la estimación del precio en términos absolutos (€/mes) y unitarios (€/m²), e incluye indicadores de variación del precio estimado en el último trimestre y en el último año, con el fin de contextualizar el componente temporal del mercado. Para mejorar la interpretabilidad, se añade una explicación textual breve que resume los factores más relevantes y una gráfica de evolución histórica del precio estimado, permitiendo situar la predicción dentro de una trayectoria temporal comparable.



(a) **(b)**

Figura 5. Pantallas de la aplicación web**: (a)** Características del inmueble y localización geográfica. **(b)** Resultados de la estimación del precio de alquiler

4. CONCLUSIONES

Este trabajo ha permitido diseñar, optimizar e interpretar un conjunto de modelos de aprendizaje automático orientados a la estimación del precio de oferta del alquiler residencial para la Comunidad Valenciana. A partir de un proceso estructurado de entrenamiento, validación y evaluación sobre datos

georreferenciados, se constata que los métodos *ensemble* basados en *boosting* —en particular LightGBM (LGBM) y Gradient Boosting Regressor (GBR)— ofrecen el mejor equilibrio entre precisión predictiva, capacidad de generalización y eficiencia computacional en el contexto analizado.

La optimización de hiperparámetros mediante validación cruzada ha mostrado que la combinación de búsqueda aleatoria y optimización bayesiana es útil para explorar el espacio de configuraciones, aunque las diferencias en el desempeño final entre ambas estrategias resultan limitadas.

Desde la perspectiva interpretativa, la combinación de importancia por permutación, gráficos de dependencia parcial (PDP) y explicaciones locales mediante SHAP permite extraer conclusiones consistentes sobre los determinantes del alquiler ofertado. Las variables con mayor contribución son la ubicación geográfica y las características estructurales de la vivienda como superficie construida y número de baños. En cambio, atributos como la presencia de ascensor muestran un impacto marginal en el mercado de alquiler del ámbito estudiado, al menos en comparación con los factores espaciales y los atributos principales del inmueble.

En términos aplicados, el estudio culmina con el despliegue de una aplicación web de acceso abierto que permite a usuarios particulares, profesionales e inversores obtener una estimación del precio actual (y su contextualización temporal) a partir de las características del inmueble y su localización. La incorporación de explicaciones basadas en SHAP aporta trazabilidad al resultado y mejora la comprensibilidad del modelo, un aspecto especialmente relevante cuando se pretende democratizar el uso de herramientas predictivas en entornos reales.

5. ABREVIATURAS Y ACRÓNIMOS

GBR	Gradient Boosting Regressor
XGBM	eXtreme Gradient Boosting
LGBM	Light Gradient Boosting Machine (o LightGBM)
RF	Random Forest
ET	Extra Trees Regressor
MAE	Mean Absolute Error (Error Absoluto Medio)
MSE	Mean Square Error (Error Cuadrático Medio)
RMSE	Root Mean Square Error (Raíz del Error Cuadrático Medio)
R^2	Coeficiente de Determinación R^2
SHAP	Valores de Shapley
IA	Inteligencia Artificial

XAI	eXplainable Artificial Intelligence (Inteligencia Artificial Explicable)
ML	Machine Learning (Aprendizaje Automático)
PDP	Partial Dependence Plot (Gráfico de Dependencia Parcial)

6. FINANCIACIÓN

Proyecto de investigación con referencia GRE23-05A, “Ayudas para proyectos y redes de investigación, Categoría A (Anexo X)”, financiado por la Universidad de Alicante (BOUA 10/05/2023). Se desarrolló un prototipo inicial del proyecto en el marco del Proyecto de Investigación con referencia CENID2024/12, financiado por el Centro de Inteligencia Digital de la Universidad de Alicante (CENID).

7. BIBLIOGRAFÍA

[1] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado*, et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion,* vol. 58, pp. 82-115, 2020.

[2] B. Park y J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data", *Expert Systems with Applications,* vol. 42, no. 6, pp. 2928-2934, 2015.

[3] E. A. Antipov y E. B. Pokryshevskaya, "Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics", *Expert Systems with Applications,* vol. 39, no. 2, pp. 1772-1778, 2012.

[4] M. Čeh, M. Kilibarda, A. Lisec, y B. Bajat, "Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments", *ISPRS International Journal of Geo-Information,* vol. 7, no. 5, pp. 168, 2018.

[5] W. T. Embaye, Y. A. Zereyesus, y B. Chen, "Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches", *PLOS ONE,* vol. 16, no. 2, pp. e0244953, 2021.

[6] S. Gnat, "Property Mass Valuation on Small Markets", *Land,* vol. 10, no. 4, pp. 388, 2021.

[7] J. Hong, "An Application of XGBoost, LightGBM, CatBoost Algorithms on House Price Appraisal System", *Housing Finance Research,* vol. 4, pp. 33-64, 2020.

[8] J. Hong, H. Choi, y W.-S. Kim, "A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea", *International Journal of Strategic Property Management,* vol. 24, no. 3, pp. 140-152, 2020.

[9] N. Kok, E.-L. Koponen, y C. A. Martínez-Barbosa, "Big Data in Real Estate? From Manual Appraisal to Automated Valuation", *Journal of Portfolio Management,* vol. 43, no. 6, pp. 202-211, 2017.

[10] J. R. Rico-Juan y P. Taltavull de La Paz, "Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain", *Expert Systems with Applications,* vol. 171, pp. 114590, 2021.

[11] L. Xu y Z. Li, "A New Appraisal Model of Second-Hand Housing Prices in China's First-Tier Cities Based on Machine Learning Algorithms", *Computational Economics,* vol. 57, no. 2, pp. 617-637, 2021.

[12] S. Yilmazer y S. Kocaman, "A mass appraisal assessment study using machine learning based on multiple regression and random forest", *Land Use Policy,* vol. 99, pp. 104889, 2020.

[13] R.-T. Mora-Garcia, M.-F. Cespedes-Lopez, y V. R. Perez-Sanchez, "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times", *Land,* vol. 11, no. 11, pp. 2100, 2022.

[14] J.-L. Alfaro-Navarro, E. L. Cano, E. Alfaro-Cortés, N. García, M. Gámez, y B. Larraz, "A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems", *Complexity,* vol. 2020, pp. 5287263, 2020.

[15] S. Canaz Sevgen y Y. Aliefendioğlu, "Mass Apprasial With A Machine Learning Algorithm: Random Forest Regression", *Bilişim Teknolojileri Dergisi,* vol. 13, no. 3, pp. 301-311, 2020.

[16] W. K. O. Ho, B.-S. Tang, y S. W. Wong, "Predicting property prices with machine learning algorithms", *Journal of Property Research,* vol. 38, no. 1, pp. 48-70, 2021.

[17] L. Hu, S. He, Z. Han, H. Xiao, S. Su, M. Weng*, et al.*, "Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies", *Land Use Policy,* vol. 82, pp. 657-673, 2019.

[18] P.-F. Pai y W.-C. Wang, "Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices", *Applied Sciences,* vol. 10, no. 17, pp. 5832, 2020.

[19] INE, Instituto Nacional de Estadística. Atlas de distribución de renta de los hogares [Online]. Available: https://www.ine.es/componentes_inebase/ADRH_total_nacional.htm